On the Choice of Prior in Bayesian Model Averaging

Einmahl, J.H.J.; Magnus, J.R.; Kumar, K.

Publication date:
2011

No. 2011-003

ON THE CHOICE OF PRIOR IN BAYESIAN MODEL AVERAGING

By John H.J. Einmahl, Kamlesh Kumar, Jan R. Magnus

January 2011

ISSN 0924-7815

On the choice of prior in Bayesian model averaging*

January 13, 2011

John H.J. Einmahl
Department of Econometrics & Operations Research and CentER, Tilburg University

Kamlesh Kumar
CentER, Tilburg University

Jan R. Magnus
Department of Econometrics & Operations Research and CentER, Tilburg University

*We are grateful to Estate Khmaladze and Hannes Leeb for helpful comments, and to Wendun Wang for computational assistance. E-mail addresses: j.h.j.einmahl@uvt.nl (John Einmahl), k.kumar@uvt.nl (Kamlesh Kumar), magnus@uvt.nl (Jan Magnus).

Corresponding author:
Jan R. Magnus
Department of Econometrics & Operations Research and CentER
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands
e-mail: magnus@uvt.nl

Abstract: Bayesian model averaging attempts to combine parameter estimation and model uncertainty in one coherent framework. The choice of prior is then critical. Within an explicit framework of ignorance we define a 'suitable' prior as one which leads to a continuous and suitable analog to the pretest estimator. The normal prior, used in standard Bayesian model averaging, is shown to be unsuitable. The Laplace (or lasso) prior is almost suitable. A suitable prior (the Subbotin prior) is proposed and its properties are investigated.

JEL Classification: C11, C51, C52.

Keywords: Model averaging, Bayesian analysis, Subbotin prior.

1 Introduction

Suppose a ruler seeks advice on a specific parameter, say next year's inflation. He has twelve advisors, and each advisor provides an estimate. When all have left, the ruler has twelve estimates. In addition, he has an opinion about each advisor based on past experience and current performance. How does the ruler now obtain a single estimate? Let us consider two possibilities. The ruler may think: Whom do I trust most? Whose advice do I think most reliable? Then, he takes the advice of his most trusted advisor. This is the first method. Alternatively, he may consider all advisors useful, but not to the same degree. Some are more experienced and more clever than others, so they get a higher weight. Then, the ruler computes a weighted average of the twelve estimates. This is the second method.

While the second method appeals to common sense, econometric practice favors the first method. In econometric practice one typically first selects a model and then computes estimates within this selected model. There are many alternative models, each of which provides some insight into the value of the required parameter, and these other models play a role in the model selection procedure, but not in the estimation procedure. Model selection and estimation are completely separated, just like the ruler only listening to his most trusted advisor. In econometrics, the first method is called 'pretesting'. The second method, called 'model averaging', not only appeals to common sense, but also has two major advantages. First, it avoids arbitrary thresholds (like 1.96), thus forcing continuity on a previously discontinuous estimator; second, it allows us to combine model selection and estimation into *one* procedure, thus moving from conditional to unconditional estimator characteristics.

Standard Bayesian model averaging (BMA), though appealing in principle, has however three drawbacks. First, the prior is almost always assumed to be normal, and this is not good, because it leads to unbounded risk and other undesirable properties caused by the fact that the tails of the normal distribution are too thin. (In this paper, risk is always defined under squared-error loss; hence it is synonymous to mean squared error.) Second, the variance in the prior distribution (such as Zellner's g-prior) is chosen ad hoc for computational ease without theoretical justification. Third, the computational effort required in applying BMA is very large, so that approximating methods like Markov chain Monte Carlo (MCMC) are needed.

In a recent paper, Magnus et al. (2010) resolved all three problems by in-

troducing the weighted-average least squares (WALS) procedure and proposing the Laplace prior. One of the features of this approach is an explicit treatment of ignorance. Bayesian methods are designed for situations where prior information is available, in which case they guide us as to how this information should be combined with information from the data. But when no prior information is available and nevertheless informative priors need to be specified (as is the case with BMA), then an explicit treatment of ignorance is required. We shall not define the vague concept of ignorance, but we shall make ignorance explicit, at least in part, by introducing the concept of 'neutrality' in Section 4.

The econometric framework is the regression model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon = X\beta + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 I_n), \tag{1}$$

where y $(n \times 1)$ is the vector of observations, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are matrices of deterministic regressors, ε is a random vector of unobservable disturbances, and β_1 and β_2 are unknown parameter vectors. We assume that $k_1 \geq 1$, $k_2 \geq 0$, $k := k_1 + k_2 \leq n - 1$, and that $X := (X_1 : X_2)$ has full column-rank. The reason for distinguishing between X_1 and X_2 is that X_1 contains explanatory variables which we want in the model on theoretical or other grounds, while X_2 contains additional explanatory variables of which we are less certain. The columns of X_1 are called 'focus' regressors, and the columns of X_2 'auxiliary' regressors.

In the simplest case $k_2 = 1$ we have only one auxiliary regressor, which we write as x_2 rather than as X_2, and only two models to consider: restricted (where $\beta_2 = 0$) and unrestricted. The least-squares estimator of β_1 in the restricted model is $\hat{\beta}_{1r} = (X_1'X_1)^{-1}X_1'y$. In the unrestricted model we have

$$\hat{\beta}_{1u} = \hat{\beta}_{1r} - \sigma\hat{\theta}q, \qquad \hat{\beta}_{2u} = \frac{x_2'M_1y}{x_2'M_1x_2},$$

where

$$M_1 := I_n - X_1(X_1'X_1)^{-1}X_1', \qquad q := \frac{1}{\sqrt{x_2'M_1x_2}}(X_1'X_1)^{-1}X_1'x_2,$$

and

$$\hat{\theta} := \frac{\hat{\beta}_{2u}}{\sigma/\sqrt{x_2'M_1x_2}} \sim \mathrm{N}(\theta, 1), \qquad \theta := \frac{\beta_2}{\sigma/\sqrt{x_2'M_1x_2}}. \tag{2}$$

Now consider the weighted-average least-squares (WALS) estimator of β_1,

$$b_1 := \lambda\hat{\beta}_{1u} + (1-\lambda)\hat{\beta}_{1r} = \hat{\beta}_{1r} - \sigma(\lambda\hat{\theta})q,$$

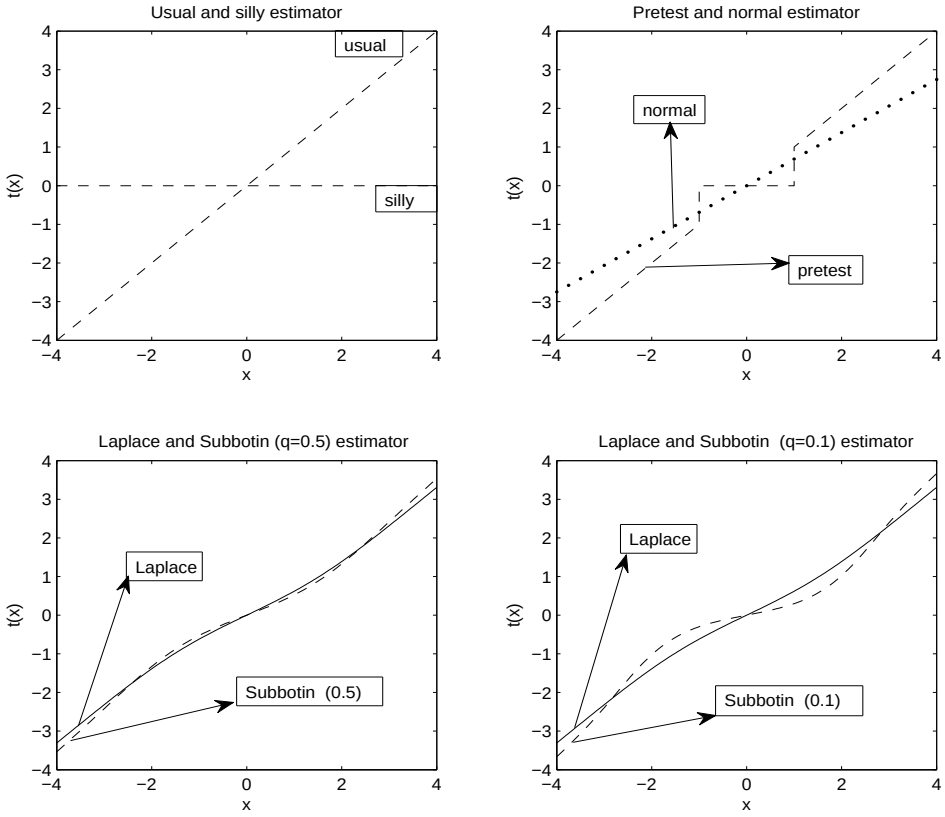

Figure 1: Estimators $t(x)$ of θ when $x \sim \mathrm{N}(\theta, 1)$.

where $0 \leq \lambda \leq 1$ may depend on y, but only through the residuals $M_1 y$. The 'equivalence theorem' (in its simplest form) now implies that

$$\mathrm{MSE}(b_1) = \sigma^2 \left((X_1' X_1)^{-1} + \mathrm{MSE}(\lambda\hat{\theta}) qq' \right), \tag{3}$$

where $\lambda\hat{\theta}$ is a (shrinkage) estimator of θ. The equivalence theorem (Magnus and Durbin, 1999, Theorem 2; Danilov and Magnus, 2004, Theorem 1) is in fact much more general, but its essence is well reflected in (3). The theorem implies that if we can find a λ-function such that $\lambda\hat{\theta}$ is an optimal estimator of θ (in the mean squared error sense), then *the same* λ-function will provide an optimal WALS estimator of β_1. The problem of estimating β_1 in a regression context is thus reduced to estimating θ from a single observation $\hat{\theta} \sim \mathrm{N}(\theta, 1)$. The idea of not analyzing the general regression problem, but only the estimation of the mean θ from a single observation x with known variance, goes back to Bickel (1983, 1984).

We study this seemingly trivial problem in isolation, writing x instead of $\hat{\theta}$. Thus, we have one observation $x|\theta \sim \mathrm{N}(\theta, 1)$, and the question is how to estimate θ. In Figure 1 we consider seven estimators $t(x)$ of θ. First, the 'usual' estimator $t(x) = x$ and the 'silly' estimator $t(x) = 0$. Then the pretest estimator, a combination of the 'usual' and 'silly' estimators:

$$t(x) = \begin{cases} x & \text{if } |x| > c, \\ 0 & \text{if } |x| \leq c, \end{cases}$$

for some $c > 0$. Next four Bayesian estimators, based on a prior π on θ. These can be viewed as continuous versions of the pretest estimator: the 'normal' estimator based on a normal prior, the Laplace estimator based on a Laplace prior, and two Subbotin estimators based on a Subbotin prior (introduced in Section 4).

The 'usual' estimator is good when x is large but not when x is small, while the 'silly' estimator is good when x is small, but not when x is large. The pretest estimator is a compromise, but not a very good compromise, since it is discontinuous and hence inadmissable. The 'normal' estimator has unbounded risk and $x - t(x)$ diverges to ∞ as $x \to \infty$. The Laplace estimator has bounded risk and $x - t(x)$ converges to a positive constant, but not to zero. This is undesirable, because we would wish that, for large x, the estimator 'is close to x' in the sense that $x - t(x)$ converges to zero. The Subbotin estimator, which is the one advocated in this paper, possesses this property. This estimator depends on a parameter $0 < q < 1$ and Figure 1 shows (for $q = 0.5$ and $q = 0.1$) that the Subbotin estimator curves back to the 45° line, in contrast to the Laplace estimator.

Bayesian averaging of Bayesian estimators was first proposed by Leamer (1978), and Bayesian averaging of classical estimators by Raftery (1995). A large literature on Bayesian model averaging now exists. An alternative approach, first proposed by Frank and Friedman (1993), is to consider penalized regression by computing

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left((y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j|^q \right)$$

for some $\lambda > 0$ and $q > 0$. The case $q = 2$ is the familiar ridge regression. Tibshirani (1996) studied the case $q = 1$, which he called the lasso (least absolute shrinkage and selection operator). Park and Casella (2008) and Hans (2009) exploited the fact that the lasso has a Bayesian counterpart, which they called the lasso prior. The lasso prior is in fact the Laplace density. The Subbotin estimator proposed in the current paper is related

to the penalized regression problem when $0 < q < 1$. A different penalty function called SCAD (smoothly clipped absolute deviation) was introduced by Fan and Li (2001). Our proposed (Bayesian) estimator is inspired by the SCAD (non-Bayesian) estimator in that it 'curves back' to the 45° line when x is large, as is well illustrated in Figure 2 of Fan and Li (2001). What they call the 'hard' thresholding function is in fact the pretest estimator, and their lasso and SCAD thresholding functions in a non-Bayesian context are discontinuous counterparts of the Laplace and Subbotin estimators in a Bayesian context.

The purpose of this paper is thus to find a suitable prior π, which appropriately reflects our notion of ignorance and, when combined with a single observation $x|\theta \sim \mathrm{N}(\theta, 1)$, leads to an estimator $t(x)$ satisfying all properties which we consider to be desirable. In Section 2 we impose some general restrictions on the prior π (Assumption A) and obtain some important properties of the posterior mean $t(x)$ and of the posterior variance $\sigma^2(x)$, while assuming only that the prior π satisfies Assumption A. In Section 3 we note that Assumption A does not imply that '$t(x)$ is close to x' when x is large. Our main result is Theorem 1, where we provide necessary and sufficient conditions on the prior such that $x - t(x) \to 0$ as $x \to \infty$. In search of a prior which satisfies Assumption A and the condition of Theorem 1, we arrive in Section 4 at the Subbotin prior

$$\pi(\theta) = \frac{q c^{1/q}}{2\Gamma(1/q)}\, e^{-c|\theta|^q} \qquad (c > 0,\, q > 0),$$

and we prove in Theorem 2 that this prior is 'suitable' if and only if $0 < q < 1$. We obtain properties of the implied estimator $t(x)$, the posterior variance $\sigma^2(x)$, the deviation $x - t(x)$, and the shrinkage factor $t(x)/x$. To gain further insight into the properties of the estimator $t(x)$ based on the Subbotin prior, we take a frequentist view in Section 5, where we consider the bias, standard deviation, and root mean squared error of $t(x)$ as an estimator of θ, when x is an observation from the $\mathrm{N}(\theta, 1)$ distribution. Then, in Section 6, we move to the regression context and to WALS estimation, and reconsider two growth models typically estimated in the growth empirics literature. Our main aim in this section is to analyze the sensitivity of our results to the choice of prior. Section 7 concludes. Appendix A gives some background mathematical results, and Appendix B provides the proofs of the two theorems.

2 Properties of the posterior mean

Throughout we shall impose the following restrictions on the prior density.

Assumption A: The prior π is

(A.1) symmetric around zero: $\pi(-\theta) = \pi(\theta)$ for all $\theta > 0$;

(A.2) positive and non-increasing on $(0, \infty)$;

(A.3) differentiable, except possibly at 0;

and

(A.4) $\omega(\theta) := -\pi'(\theta)/\pi(\theta)$ has a limit (possibly ∞) as $\theta \to \infty$.

Letting ϕ denote the standard-normal density, we can write the posterior density as

$$p(\theta|x) = \frac{\phi(x-\theta)\pi(\theta)}{\int_{-\infty}^{\infty} \phi(x-\theta)\pi(\theta)\, d\theta}.$$

Define, for $j = 0, 1, \ldots,$

$$A_j(x) := \int_{-\infty}^{\infty} (x-\theta)^j \phi(x-\theta)\pi(\theta)\, d\theta.$$

Then, $\mu_j(x) := \mathrm{E}((x-\theta)^j|x) = A_j(x)/A_0(x)$, and hence the mean and variance of θ in the posterior distribution are:

$$t(x) = -\mu_1(x) + x, \qquad \sigma^2(x) = \mu_2(x) - \mu_1^2(x). \tag{4}$$

In this section we study some features of the posterior density $p(\theta|x)$, while assuming only that the prior π satisfies Assumption A. This allows us to obtain some important properties of the posterior mean $t(x)$, the Bayes estimator under squared-error loss, and of the posterior variance $\sigma^2(x)$.

In addition to the mean and variance we obtain the skewness

$$\gamma_1(x) := \frac{\mathrm{E}((\theta - t(x))^3|x)}{\sigma^3(x)} = \frac{-\mu_3(x) + 3\mu_1(x)\mu_2(x) - 2\mu_1^3(x)}{(\mu_2(x) - \mu_1^2(x))^{3/2}},$$

and the kurtosis

$$\begin{aligned}\gamma_2(x) :&= \frac{\mathrm{E}((\theta - t(x))^4|x)}{\sigma^4(x)} - 3 \\ &= \frac{\mu_4(x) - 3\mu_2^2(x) - 4\mu_1(x)\mu_3(x) + 12\mu_1^2(x)\mu_2(x) - 6\mu_1^4(x)}{(\mu_2(x) - \mu_1^2(x))^2}.\end{aligned}$$

Since $A_0'(x) = -A_1(x)$ and

$$A_j'(x) = jA_{j-1}(x) - A_{j+1}(x) \qquad (j = 1, 2, \ldots),$$

we obtain

$$\mu_j'(x) = j\mu_{j-1}(x) - \mu_{j+1}(x) + \mu_1(x)\mu_j(x) \qquad (j = 1, 2, \dots),$$

and hence,

$$t'(x) = \sigma^2(x), \qquad t''(x) = \sigma^3(x)\gamma_1(x), \qquad t'''(x) = \sigma^4(x)\gamma_2(x). \tag{5}$$

The posterior density is non-degenerate for any x. Hence, $t'(x) = \sigma^2(x) > 0$ for any x, and t is increasing on $(-\infty, \infty)$. The function t is odd: $t(-x) = -t(x)$, so that $t(0) = 0$. Also,

$$0 < t(x) < x \qquad (x > 0). \tag{6}$$

To prove (6) we need two results from Appendix A. It is clear that $A_0(x) > 0$ for all x. It follows from (14) that $A_1(0) = 0$ and that $A_1(x) > 0$ for all $x > 0$. Hence, $\mu_1(x) \geq 0$ for all $x \geq 0$ with equality if and only if $x = 0$. It follows from (15) that $\mu_1(x) \leq x$ with equality if and only if $x = 0$. Hence, $\mu_1(0) = 0$ and $0 < \mu_1(x) < x$ for all $x > 0$. Equation (6) then follows from (4).

The derivatives in (5) show in particular that, at $x = 0$,

$$t(0) = 0, \qquad t'(0) = \mu_2(0) > 0, \qquad t''(0) = 0, \qquad t'''(0) = \mu_2^2(0)\gamma_2(0).$$

In fact, all even-ordered derivatives (t, t'', and so on) vanish at $x = 0$.

We already know that $\mu_2(0) > 0$. To prove that $\mu_2(0) < 1$ we note that $(1 - \theta^2)(\pi(\theta) - \pi(1)) \geq 0$. This gives

$$\int_{-\infty}^{\infty} (1 - \theta^2)\phi(\theta)\pi(\theta)\, d\theta > \pi(1) \int_{-\infty}^{\infty} (1 - \theta^2)\phi(\theta)\, d\theta = 0.$$

The inequality is strict, because $(1 - \theta^2)(\pi(\theta) - \pi(1)) = 0$ holds only for all θ if π is constant and hence improper. Thus we have proved

$$0 < \mu_2(0) < 1.$$

We know that t is increasing. We now show that, if Assumption A holds, then

$$t(x) \to \infty \quad \text{as} \quad x \to \infty.$$

Let $M > 1$. Then, using (4), and (9), (12), and (15) from Appendix A, we obtain for all $x > 1$:

$$t(x) = \frac{\int_0^\infty \theta \left(e^{\theta x} - e^{-\theta x}\right) \phi(\theta)\pi(\theta)\, d\theta}{\int_0^\infty \left(e^{\theta x} + e^{-\theta x}\right) \phi(\theta)\pi(\theta)\, d\theta} \geq \frac{M}{4} \, \frac{\int_M^\infty e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta}{\int_0^\infty e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta}.$$

Also, the limiting result (18) in Appendix B implies

$$\int_0^M e^{\theta x}\phi(\theta)\pi(\theta)\,d\theta \leq \int_M^\infty e^{\theta x}\phi(\theta)\pi(\theta)\,d\theta$$

for x sufficiently large. The two inequalities together show that $t(x) \geq M/8$, and since M can be arbitrarily large, t is unbounded.

The variance function σ^2 is even: $\sigma^2(-x) = \sigma^2(x)$, and satisfies $\sigma^2(x) > 0$ for all x, and $\sigma^2(0) = \mu_2(0)$. For small x we have

$$\sigma^2(x) = \mu_2(0) + \frac{1}{2}x^2\mu_2^2(0)\gamma_2(0) + O(x^4),$$

so that σ^2 is increasing on a small interval $(0, \delta)$ if the kurtosis $\gamma_2(0) > 0$.

Finally, we consider the 'shrinkage' function $\lambda(x) := t(x)/x$. This function is also even: $\lambda(-x) = \lambda(x)$, and it satisfies, because of (6),

$$0 < \lambda(x) < 1 \qquad (x > 0).$$

From the expansion of $t(x)$ we have, for small x,

$$\lambda(x) = \mu_2(0) + \frac{1}{6}x^2\mu_2^2(0)\gamma_2(0) + O(x^4),$$

so that λ (like σ^2) is increasing on a small interval $(0, \delta)$ if $\gamma_2(0) > 0$.

3 Asymptotic behavior of $x - t(x)$

The above analysis shows that a number of desirable properties of $t(x)$ follow from Assumption A. The assumption does not, however, imply that '$t(x)$ is close to x' when x is large. This is important because, if we have one observation x from the $\mathrm{N}(\theta, 1)$ distribution, then the 'usual' estimator may be criticized for not performing well near zero, but for large x it is the obvious candidate. To make the statement '$t(x)$ is close to x' precise, we need the following definition.

Definition 1 (suitability): *Consider a univariate random variable x with $x|\theta \sim \mathrm{N}(\theta, 1)$ together with a prior π on θ. Let $t(x)$ and $\sigma^2(x)$ denote the mean and variance of θ in the posterior distribution. If π satisfies Assumption A, and $x - t(x) \to 0$ and $\sigma^2(x) \to 1$ as $x \to \infty$, then π is called suitable.*

We note that if π is suitable, then a third requirement, namely that the

shrinkage factor $\lambda(x) \to 1$ as $x \to \infty$, is trivially fulfilled. In fact, we have $\lambda(x) = 1 + o(1/x)$.

To see why Assumption A is not sufficient for any of these three requirements to hold, consider the prior $\pi(\theta) \propto \exp(-c|\theta|^q)$, where $c > 0$ and $q > 0$. (The symbol $\propto$ means: is proportional to.) For $q > 2$ (thin tails), none of the three properties holds: $\lambda(x)$ and $\sigma^2(x)$ converge to zero, and $x - t(x)$ converges to infinity. For $q = 2$ (normal prior), $\lambda(x) < 1$ and $\sigma^2(x) < 1$ are constant, and $x - t(x)$ goes to infinity. For $q = 1$ (Laplace prior), two of the three properties hold. Only $x - t(x)$ goes to a constant, not zero. This suggests that suitability occurs when $0 < q < 1$. We shall see in Section 4 that this is indeed the case.

But first we consider the limiting behavior of $g(x) := x - t(x)$ for a general prior π satisfying Assumption A. Define the hazard (or failure) rate as

$$h(\theta) := \frac{\pi(\theta)}{\int_\theta^\infty \pi(\xi)\, d\xi}.$$

The hazard rate is a well-established characteristic of a distribution. Hazard rates can be increasing or decreasing (or both). If somebody has lived a long time, he or she will probably die soon (positive duration dependence, increasing hazard rate). But someone without employment has a higher chance of finding a job the shorter is the unemployment spell (negative duration dependence, decreasing hazard rate).

Our main result gives a complete characterization, in terms of the hazard rate, of the class of priors π with the property that $g(x) \to 0$ as $x \to \infty$, and can be stated as follows.

Theorem 1: *Under Assumption A,*

$$\lim_{\theta\to\infty} h(\theta) = 0 \iff \lim_{x\to\infty} g(x) = 0.$$

The tail of π can be classified according to the limiting behavior of h. The condition $\lim_{\theta\to\infty} h(\theta) = 0$ means that the tail of π is 'medium-long' or 'long'; see Schuster (1984).

4 The Subbotin prior

In search of a suitable prior, let us consider a general class of densities, the reflected generalized gamma density,

$$\pi(\theta) \propto |\theta|^{-\alpha} e^{-c|\theta|^q} \qquad (c > 0,\ q > 0,\ \alpha < 1).$$

For $q = 1$ this becomes the reflected gamma density, and other special cases are the reflected Weibull ($\alpha + q = 1$), Laplace ($\alpha = 0$, $q = 1$), and normal ($\alpha = 0$, $q = 2$) densities.

To satisfy Assumption A we must have $0 \leq \alpha < 1$. If $\alpha > 0$ then $\pi(0) = \infty$. Since we wish to stay close to the Laplace prior (where $\pi(0)$ is finite), our interest is in the special case $\alpha = 0$:

$$\pi(\theta) = \frac{qc^{1/q}}{2\Gamma(1/q)}\, e^{-c|\theta|^q} \qquad (c > 0,\, q > 0). \tag{7}$$

This density was first proposed by Subbotin (1923), and we shall call it the 'Subbotin' density. Special cases are the normal density ($q = 2$) and the Laplace density ($q = 1$). To satisfy the condition of Theorem 1 we must have $0 < q < 1$, and in fact this is precisely what is required for suitability.

Theorem 2: *The Subbotin prior is suitable if and only if* $0 < q < 1$.

All moments of the Subbotin density exist. Since the density is symmetric around zero, the odd moments vanish. The even moments are given by

$$\mathrm{E}(\theta^{2m}) = \frac{\Gamma((2m+1)/q)/\Gamma(1/q)}{c^{2m/q}}.$$

In particular, the variance and kurtosis of θ are

$$\mathrm{var}(\theta) = \frac{\Gamma(3/q)/\Gamma(1/q)}{c^{2/q}}, \qquad \mathrm{kur}(\theta) = \frac{\Gamma(5/q)/\Gamma(3/q)}{\Gamma(3/q)/\Gamma(1/q)} - 3.$$

The Subbotin prior involves two parameters: c and q. How should these parameters be chosen? As discussed in the Introduction, we wish to make ignorance explicit, at least in part. The quantification of ignorance in a Bayesian setting is an old problem and was already discussed in the classical essays of Bayes (1763) and Laplace (1814); see Bauwens et al. (1999, Section 4.3), who also discuss, among others, Jeffreys' (1961) invariance principle and Bernardo's (1979) reference prior. Our approach to this problem is a little different. Suppose that our interest is in estimating β_1 in the model (1),

$$y = X_1\beta_1 + \beta_2 x_2 + \varepsilon,$$

where we have again taken $k_2 = 1$ for simplicity. We are not sure whether the auxiliary parameter β_2 is zero or not, so that we consider two models: the restricted (with $\beta_2 = 0$) and the unrestricted. If we include the auxiliary variable x_2 in our regression equation, then $\bar{R}^2$ (the adjusted R^2) will increase

if, and only if, the t-ratio of the auxiliary parameter is larger than one in absolute value. Also, if we define the 'theoretical' t-ratio,

$$\theta := \frac{\beta_2}{\sigma/\sqrt{x_2' M_1 x_2}},$$

as in (2), then $\text{MSE}(\hat{\beta}_{1r}) \leq \text{MSE}(\hat{\beta}_{1u})$ if, and only if, $|\theta| \leq 1$, where $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ denote the restricted and unrestricted estimator of β_1 respectively; see Magnus and Durbin (1999, Theorem 1). Based on these facts we shall say that we are 'neutral' about the auxiliary parameter β_2 when (a) we don't know whether β_2 is positive or negative, and (b) we don't know whether including the corresponding auxiliary regressor x_2 will increase or decrease the mean squared error of the estimated focus parameter β_1. More formally, we choose the prior distribution such that the prior median of θ is zero and the prior median of $|\theta|$ is one.

Definition 2 (neutrality): *Let π be a prior density. If*

$$\int_{-\infty}^{-1} \pi(\theta)\, d\theta = \int_{-1}^{0} \pi(\theta)\, d\theta = \int_{0}^{1} \pi(\theta)\, d\theta = \int_{1}^{\infty} \pi(\theta)\, d\theta = \frac{1}{4},$$

then π is called neutral.

The concept of neutrality, introduced by Magnus (2000), thus attempts to capture the vague concept of ignorance in an explicit, transparent, and applicable definition.

For a neutral Subbotin prior c is determined by q as given in Table 1 for selected values of q. Under neutrality, both the standard deviation (sd) and kurtosis (kur) of the prior increase when q decreases, because the tail becomes thicker. We also present two quantiles. Neutrality implies that $Q(4/8) = 0$ and $Q(6/8) = 1$. The table gives $Q(5/8)$ and $Q(7/8)$. If the prior is relatively flat between 0 and 1, then $Q(5/8)$ should be close to 0.5. We see that the smaller is q, the less flat is the prior between 0 and 1. Small values of q (like $q = 0.1$) are therefore unappealing from the point of view of ignorance. We could partially repair this by modifying the Subbotin prior, for example by introducing

$$\pi(\theta) \propto e^{-c(1-q+|\theta|)^q},$$

which we call the 'modified' Subbotin prior (Subbotin-mod). For $q = 1$ we get the Laplace prior back, but for $0 < q < 1$ we obtain a flatter distribution between 0 and 1. In Table 1 we present the corresponding moments and

Table 1: Moments and quantiles of some neutral priors.

	q	c	sd	kur	$Q(5/8)$	$Q(7/8)$
Normal	2.0	0.2275	1.48	0.00	0.47	1.71
Laplace	1.0	0.6931	2.04	3.00	0.42	2.00
Subbotin	0.9	0.8011	2.19	4.03	0.40	2.06
Subbotin	0.8	0.9369	2.40	5.57	0.39	2.14
Subbotin	0.7	1.1125	2.69	8.06	0.37	2.25
Subbotin	0.6	1.3478	3.14	12.6	0.35	2.38
Subbotin	0.5	1.6783	3.89	22.2	0.33	2.57
Subbotin	0.4	2.1757	5.37	49.0	0.30	2.86
Subbotin	0.3	3.0066	9.21	171	0.25	3.35
Subbotin	0.2	4.6709	27.1	1956	0.20	4.37
Subbotin	0.1	9.6687	691	2,823,513	0.11	8.07
Subbotin-mod	0.5	2.0323	3.16	16.3	0.37	2.33
Subbotin-mod	0.1	17.0367	14.0	85,488	0.34	2.78

quantiles for $q = 0.5$ and $q = 0.1$. For $q = 0.1$ the difference between Subbotin and modified Subbotin is large, but for $q = 0.5$ it is small.

In practical applications we recommend the Subbotin prior with $q = 0.5$. The prior then takes the form

$$\pi(\theta) = \frac{c^2}{4} e^{-c|\theta|^{1/2}}. \tag{8}$$

Since $1/q = 2$ is an integer, we can write explicitly, for $\theta > 0$,

$$\int_0^\theta \pi(\xi)\, d\xi = \frac{1}{2}\left(1 - e^{-c\theta^{1/2}}(1 + c\theta^{1/2})\right).$$

Hence, for neutrality we find c from the equation $e^c = 2(c + 1)$.

Combining the prior with the observation $x|\theta \sim \mathrm{N}(\theta, 1)$, we obtain characteristics of the posterior distribution. In Figure 1 we presented the function t. In Figures 2–4 we present the functions σ^2, g, and λ for four priors: normal ($q = 2$), Laplace ($q = 1$), and Subbotin ($q = 0.5$ and $q = 0.1$). For the normal prior, the posterior variance is constant at $\sigma^2 = 1/(2c+1) \approx 0.6873$. For the Laplace prior the variance increases to one, and for the Subbotin prior with $q < 1$, the variance also tends to one, but $\sigma^2(x)$ is not an increasing function. Regarding the deviations $g(x)$, we have $t(x) = \lambda x$ with constant $\lambda < 1$ for the normal prior, and hence $g(x) = (1 - \lambda)x$, which diverges to infinity. For the Laplace prior we have $g(x) \to \log 2 \approx 0.6931$, and for the Subbotin prior with $q < 1$ we have $g(x) \to 0$. Finally, the shrinkage factor

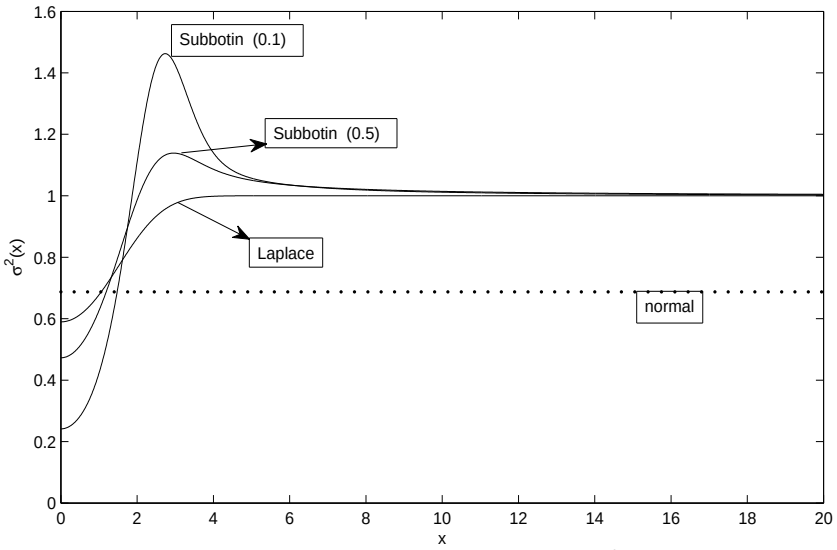

Figure 2: Posterior variances $\sigma^2(x)$.

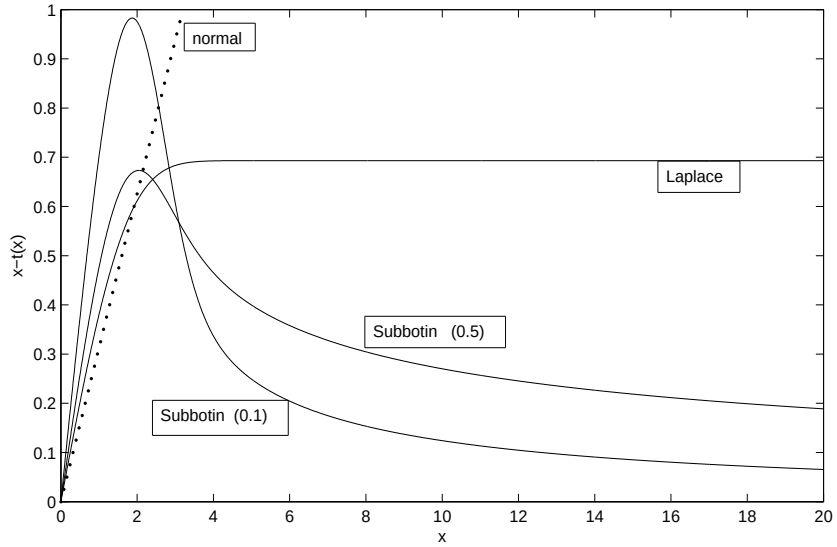

Figure 3: Deviations $g(x) = x - t(x)$.

$\lambda(x) = 1/(2c+1) \approx 0.6873$ is constant for the normal prior, and therefore does not converge to one. For $q \leq 1$, $\lambda(x) < 1$ converges to one, and the graphs in Figure 4 suggest that λ increases.

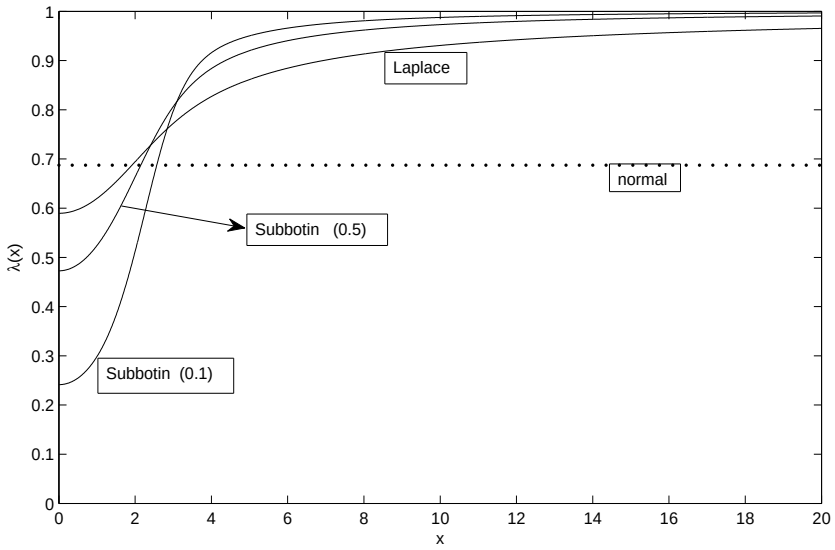

Figure 4: Shrinkage factors $\lambda(x) = t(x)/x$.

5 Root mean squared error comparisons

So far we have considered a Bayesian framework where we have one observation $x|\theta \sim \mathrm{N}(\theta, 1)$ and a prior $\pi(\theta)$. This leads to a posterior distribution and a posterior mean $t(x)$, the Bayes estimator. In this section we take a frequentist point of view and consider the properties of $t(x)$ as an estimator of θ when $x \sim \mathrm{N}(\theta, 1)$. In this framework θ is an unknown deterministic parameter. Thus we are interested in the bias $\mathrm{E}(t(x) - \theta)$, and the standard deviation (SD) and the root mean squared error (RMSE),

$$\mathrm{SD} = \sqrt{\mathrm{E}(t(x) - \mathrm{E}(t(x)))^2}, \qquad \mathrm{RMSE} = \sqrt{\mathrm{E}(t(x) - \theta)^2},$$

as a function of θ. We consider six estimators: the 'usual' estimator $t(x) = x$, the 'silly' estimator $t(x) = 0$, and four Bayesian estimators based on a Subbotin prior, namely normal ($q = 2$), Laplace ($q = 1$), and Subbotin ($q = 0.5$ and $q = 0.1$).

The bias of the six estimators is graphed in Figure 5. The bias is non-positive when $\theta > 0$ and non-negative when $\theta < 0$. The 'usual' estimator is of course the best; its bias is zero. The 'silly' estimator is the worst. The estimator based on a normal prior is also bad: its bias diverges as $\theta \to \infty$. The bias of the Laplace estimator converges to a constant, while the bias of the Subbotin estimators for $0 < q < 1$ converges to zero. Of these two estimators, the Subbotin (0.5) estimator has a smaller bias when θ is small, while the Subbotin (0.1) estimator has a smaller bias when θ is large.

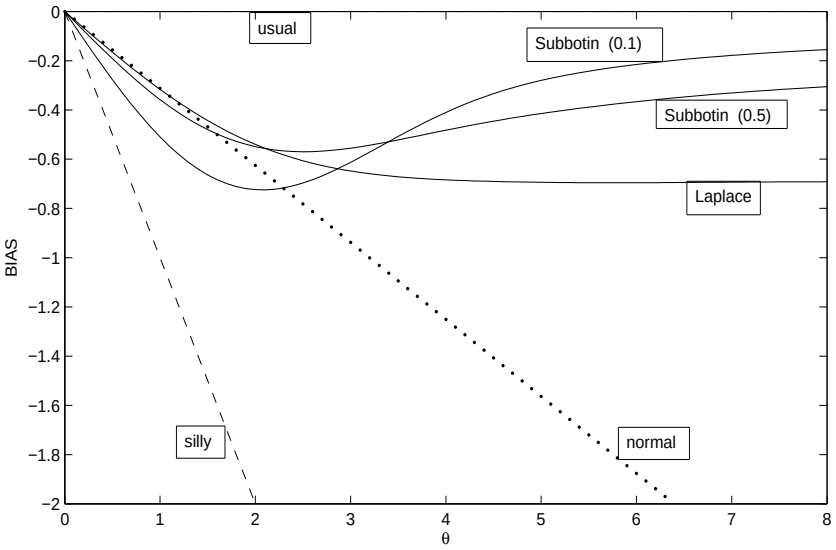

Figure 5: Bias comparisons as a function of θ.

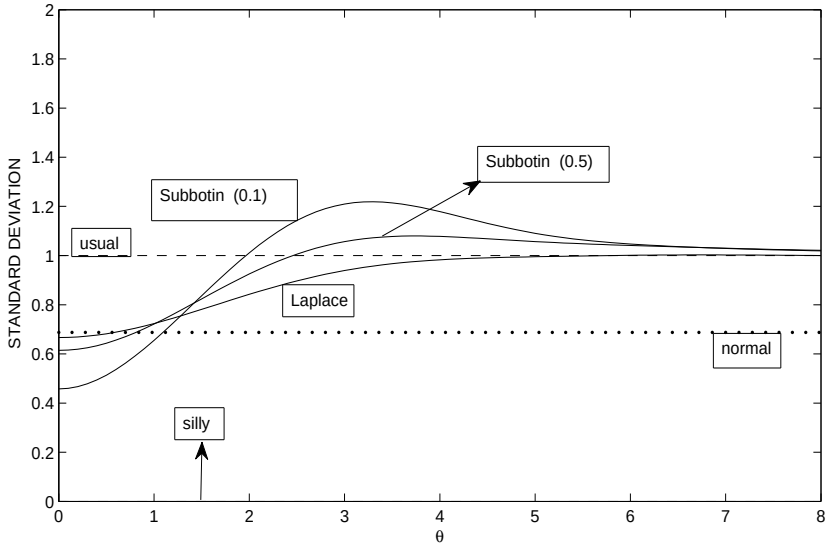

Figure 6: Standard deviation comparisons as a function of θ.

The standard deviation of the six estimators is given in Figure 6. The 'silly' estimator is of course the best; its standard deviation is zero. The standard deviation of the 'usual' estimator is one (constant). The Laplace and 'normal' estimators have standard deviation below one, while the standard deviation of the Subbotin estimators for $0 < q < 1$ is below one for small values of θ and above one for large values of θ. The standard deviation of

the 'silly' and 'normal' estimators does not converge to one.

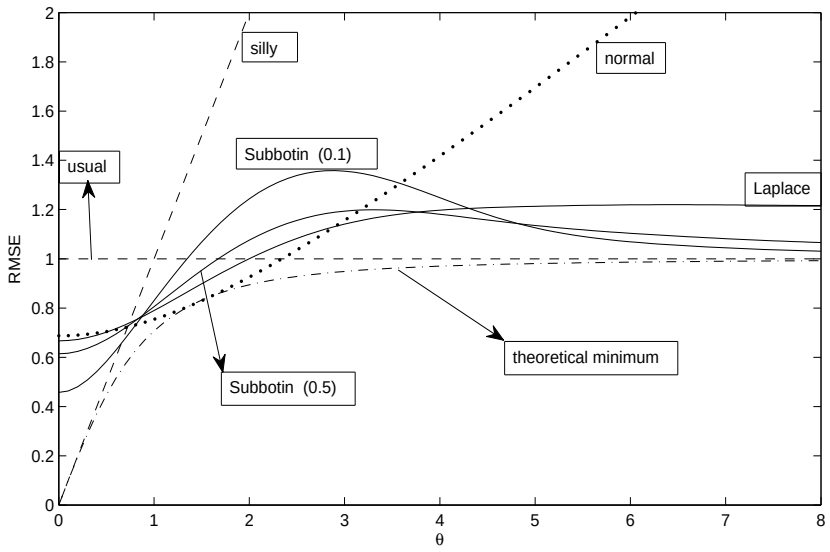

Figure 7: RMSE comparisons as a function of θ.

Most important is the combination of bias and standard deviation: the root mean squared error. This function is graphed in Figure 7 for the six estimators and for the 'theoretical minimum'. If we define $\mathcal{T}$ as the class of all real-valued functions t defined on $\mathbb{R}$ and satisfying: (a) $0 \leq t(x) \leq x$ for $x \geq 0$, (b) $t(-x) = -t(x)$ for all x, and (c) $t(x)/x$ is non-decreasing on $[0, \infty)$, then Theorem A.7 in Magnus (2002) states that

$$\inf_{t \in \mathcal{T}} \mathrm{E}\left(t(x) - \theta\right)^2 = \frac{\theta^2}{1+\theta^2}.$$

The square root of this is the 'theoretical minimum' graphed in Figure 7.

Comparing the MSE with the theoretical minimum in $\mathcal{T}$ gives us the regret,

$$\mathrm{regret}(\theta) = \mathrm{MSE}(\theta) - \frac{\theta^2}{1+\theta^2},$$

arguably a more relevant concept than risk. The regret of the 'silly' and 'normal' estimators is unbounded. The regret of the 'usual' estimator in bounded, but large for small values of θ. The Laplace estimator is 'near-optimal' in terms of regret (Magnus, 2002), but its regret does not vanish for $\theta \to \infty$. This leaves the Subbotin estimators for $0 < q < 1$. The closer we stay to the Laplace estimator (where $q = 1$) the better will be the regret performance. Hence, $q = 0.5$ is preferred over $q = 0.1$.

We conclude from the pictures that the Subbotin (0.5) estimator remains close to the Laplace estimator, but that it has the pleasant property that its bias goes to zero and its RMSE goes to one as $\theta \to \infty$, in contrast to the Laplace estimator.

6 Application to growth empirics

In order to assess the practical application of WALS estimation with the Subbotin prior, we re-estimate the models considered in Magnus et al. (2010), where WALS estimation with a Laplace prior was applied to growth empirics. We analyze two models. Model 1 can be interpreted as a direct test of the proximate neoclassical growth theory against the fundamental new growth theories of institutions, geography, fractionalization, and religion. Model 2 deviates from the proximate versus fundamental classification, and tests the robustness of the endogenous growth model using the distinction between focus and auxiliary regressors.

Table 2: Model specifications, focus and auxiliary regressors.

	Model 1	Model 2	Mean	SE
CONSTANT	Focus	Focus	1.0000	0.0000
GDP60	Focus	Focus	7.5253	0.8612
EQUIPINV	Focus	Focus	0.0432	0.0344
SCHOOL60	Focus	Focus	0.7807	0.2556
LIFE60	Focus	Focus	56.0676	1.1566
DPOP	Focus	Auxiliary	0.0206	0.0100
LAW	Auxiliary	Focus	0.5518	0.3332
TROPICS	Auxiliary	Focus	0.5481	0.4709
AVELF	Auxiliary	Focus	0.2984	0.2797
CONFUC	Auxiliary	Focus	0.0185	0.0862
MINING	—	Auxiliary	0.0482	0.0792
PRIGHTS	—	Auxiliary	3.4551	1.9073
MALARIA	—	Auxiliary	0.2866	0.4036

In both models the dependent variable is GROWTH. The regressors and their role as either focus or auxiliary are given in Table 2. Model 1 contains six focus regressors (including the constant term) and four auxiliary regressors. It is motivated by the neoclassical growth model and thus contains all Solow determinants as focus regressors, namely the constant term; the initial capital stock of an economy (GDP60), measured as the log of GDP per capita in 1960; the 1960–1985 equipment investment share of GDP (EQUIPINV); the

total gross enrollment rate in primary schooling in 1960 (SCHOOL60); life expectancy at age zero, measured in 1960 (LIFE60); and population growth rate between 1960 and 1990 (DPOP).

To test this neoclassical model and its proximate growth determinants we include four auxiliary regressors in Model 1: a rule of law index (LAW) as a measure of the importance of institutions; the country's fraction of tropical area (TROPICS); an average index of ethnolinguistic fragmentation (AVELF); and the fraction of Confucian population in the country (CONFUC).

Model 2 contains nine focus regressors and four auxiliary regressors, and it attempts more specifically to identify the factors driving growth and technological progress. All regressors of our first model are included in Model 2 as focus regressors, except DPOP which is now an auxiliary regressor. The three new auxiliary regressors are: the fraction of GDP produced in mining (MINING); an index for political rights (PRIGHTS), serving as a second institutional variable (the other is LAW); and malaria prevalence in 1966 (MALARIA), another geographical variable (next to TROPICS). A detailed explanation of all variables is given in the Appendix of Magnus et al. (2010).

For each of the two models we consider three Bayesian model averaging methods: Bayesian model averaging based on a normal prior (BMA); Weighted-average least squares (WALS) estimation based on a Laplace prior ($q = 1$); and WALS estimation based on a Subbotin prior ($q = 0.5$). The estimation results are given in Tables 3 and 4. Note that the first two columns in Tables 3 and 4 are the same as the last two columns in Tables 2 and 4 of Magnus et al. (2010). The Matlab codes used in the estimation can be downloaded from http://center.uvt.nl/staff/magnus/wals.

All regressors have the same signs across estimation methods and model specifications. The regressors GDP60, TROPICS, and AVELF in Models 1 and 2, and the regressors MINING and MALARIA in Model 2 are negatively correlated with growth, as expected. However, contrary to expectation, the coefficient of PRIGHTS is negative in Model 2. All other coefficients are positive, indicating that higher shares of physical and human capital, stability in terms of a sufficient rule of law, and a larger fraction of Confucian population have a positive influence on economic growth. Not only the signs are the same, but also the sizes of the estimates are close over the two models.

Our main interest is in the comparison of the WALS results based on the new Subbotin prior to those based on the Laplace prior. The economic impact of all robust and important regressors does not vary much between the three estimators. Comparing the estimates and standard errors of BMA with those based on the Laplace and Subbotin (0.5) priors, two conclusions emerge. First, the WALS estimates and standard errors based on Subbotin

Table 3: Estimates (and standard errors), Model 1.

	BMA	WALS with Subbotin prior	
		$q = 1.0$	$q = 0.5$
Focus regressors			
CONSTANT	0.0492 (0.0229)	0.0594 (0.0221)	0.0585 (0.0222)
GDP60	−0.0139 (0.0035)	−0.0156 (0.0033)	−0.0156 (0.0033)
EQUIPINV	0.1644 (0.0615)	0.1555 (0.0551)	0.1498 (0.0557)
SCHOOL60	0.0160 (0.0102)	0.0175 (0.0097)	0.0176 (0.0097)
LIFE60	0.0008 (0.0004)	0.0009 (0.0004)	0.0008 (0.0004)
DPOP	0.1654 (0.2770)	0.2651 (0.2487)	0.2777 (0.2523)
Auxiliary regressors			
LAW	0.0109 (0.0093)	0.0147 (0.0065)	0.0159 (0.0068)
TROPICS	−0.0035 (0.0047)	−0.0055 (0.0037)	−0.0056 (0.0040)
AVELF	−0.0021 (0.0047)	−0.0053 (0.0048)	−0.0050 (0.0046)
CONFUC	0.0612 (0.0185)	0.0443 (0.0163)	0.0468 (0.0175)

Table 4: Estimates (and standard errors), Model 2.

	BMA	WALS with Subbotin prior	
		$q = 1.0$	$q = 0.5$
Focus regressors			
CONSTANT	0.0862 (0.0239)	0.0879 (0.0246)	0.0871 (0.0246)
GDP60	−0.0164 (0.0033)	−0.0167 (0.0033)	−0.0166 (0.0033)
EQUIPINV	0.1423 (0.0553)	0.1379 (0.0562)	0.1391 (0.0558)
CONFUC	0.0550 (0.0169)	0.0550 (0.0167)	0.0549 (0.0167)
SCHOOL60	0.0162 (0.0099)	0.0156 (0.0096)	0.0157 (0.0096)
LIFE60	0.0006 (0.0003)	0.0006 (0.0003)	0.0006 (0.0003)
LAW	0.0171 (0.0067)	0.0183 (0.0066)	0.0181 (0.0067)
TROPICS	−0.0044 (0.0041)	−0.0053 (0.0040)	−0.0051 (0.0040)
AVELF	−0.0050 (0.0062)	−0.0049 (0.0059)	−0.0049 (0.0060)
Auxiliary regressors			
MINING	−0.0003 (0.0063)	−0.0056 (0.0149)	−0.0048 (0.0134)
DPOP	0.0650 (0.1705)	0.2147 (0.2178)	0.1882 (0.2177)
PRIGHTS	−0.0002 (0.0007)	−0.0008 (0.0010)	−0.0006 (0.0010)
MALARIA	−0.0072 (0.0070)	−0.0075 (0.0050)	−0.0074 (0.0053)

(0.5) are close to those based on Laplace, while BMA is not so close. Second, the estimates and standard errors of the focus variables are much closer to the Laplace results than those of the auxiliary variables. Hence, if our primary interest is on estimating the focus variables, then the practical difference

between Laplace and Subbotin (0.5) is negligible. This is partly because in our two models the t-ratios are all of moderate size, while the difference between Laplace and Subbotin would especially show for small and large t-values.

7 Concluding remarks

The strength and beauty of Bayesian model averaging lies in the fact that it combines parameter estimation and model uncertainty in one coherent framework. It provides a natural extension of pretest estimation, but while pretest estimation is discontinuous, BMA in continuous. The tradition in BMA is to make minimal assumptions on the prior. In this paper we define a class of suitable priors, which do justice to our ideas of ignorance and lead to well-behaved estimators. The normal prior, used in standard Bayesian model averaging, is not suitable. The Laplace (or lasso) prior is almost suitable. We obtain a necessary and sufficient (and easily interpreted) condition on the prior so that the estimator $t(x)$ 'is close to x' asymptotically, in the sense that $x - t(x) \to 0$ as $x \to \infty$. This is our main theoretical result (Theorem 1). Imposing this condition to the Subbotin prior, we show that it is suitable for $0 < q < 1$. We provide both Bayesian and frequentist properties of the estimator, and an empirical application to growth empirics. The Subbotin estimator when applied in the context of weighted-average least squares (WALS) combines transparency in the treatment of ignorance with a trivial computational burden (and hence shares the attractive properties of the Laplace estimator), but it has better properties than the Laplace estimator when the t-ratios of some of the auxiliary parameters are large.

Appendix A: Some properties of A_j $(j = 0, 1, 2)$

We first write the functions A_j $(j = 0, 1, \dots)$ as

$$A_j(x) = \int_0^\infty \left((x+\theta)^j \phi(x+\theta) + (x-\theta)^j \phi(x-\theta)\right) \pi(\theta)\, d\theta.$$

We have, for $j = 0, 1, \dots,$

$$A_{2j}(x) = A_{2j}(-x), \quad A_{2j+1}(x) = -A_{2j+1}(-x).$$

In this appendix we obtain alternative expressions for A_0 and A_1, and one inequality involving A_0 and A_1. The alternative expression for A_0 is easily

obtained:

$$A_0(x) = e^{-x^2/2} \int_0^\infty \left(e^{\theta x} + e^{-\theta x}\right) \phi(\theta)\pi(\theta)\, d\theta. \tag{9}$$

To obtain an alternative expression for A_1 we note first that, since π is non-increasing on $(0, \infty)$,

$$0 < \theta\pi(\theta) \leq \int_0^\theta \pi(\xi)\, d\xi \qquad (\theta > 0).$$

The right-hand side converges to 0 for $\theta \downarrow 0$, because π is integrable. Hence,

$$\lim_{\theta\downarrow 0} \theta\pi(\theta) = 0. \tag{10}$$

Define the function

$$D(\theta, x) := \phi(x + \theta) - \phi(x - \theta),$$

with partial derivative

$$D'_\theta(\theta, x) = \frac{\partial D(\theta, x)}{\partial\theta} = -(x + \theta)\phi(x + \theta) - (x - \theta)\phi(x - \theta).$$

Integrating by parts gives

$$A_1(x) = -\int_0^\infty D'_\theta(\theta, x)\pi(\theta)\, d\theta = \int_0^\infty D(\theta, x)\pi'(\theta)\, d\theta, \tag{11}$$

where we have used the fact that

$$\begin{aligned}[D(\theta, x)\pi(\theta)]_0^\infty &= \lim_{\theta\to\infty} D(\theta, x)\pi(\theta) - \lim_{\theta\to 0} D(\theta, x)\pi(\theta) \\ &= -\lim_{\theta\to 0} D(\theta, x)\pi(\theta) = 2x\phi(x) \lim_{\theta\to 0} \theta\pi(\theta) = 0,\end{aligned}$$

because of (10) and the fact that, for $\theta \to 0$,

$$\frac{D(\theta, x)}{\theta} \to D'_\theta(0, x) = -2x\phi(x).$$

Letting

$$\eta(x) = \frac{e^x - e^{-x}}{x} \quad (x > 0) \tag{12}$$

and recalling that $\omega(\theta) = -\pi'(\theta)/\pi(\theta)$, we obtain the required alternative expression for A_1 as

$$A_1(x) = x\, e^{-x^2/2} \int_0^\infty \theta\eta(\theta x)\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta. \tag{13}$$

We see that, for $x \geq 0$,

$$A_0(x) > 0, \quad A_1(x) \geq 0, \quad A_2(x) > 0, \tag{14}$$

where the first and third inequalities are trivial, and the second inequality follows from (13). In fact, $A_1(0) = 0$ and $A_1(x) > 0$ for $x > 0$. This fact together with $A_0'(x) = -A_1(x)$ implies that A_0 is decreasing on $[0, \infty)$ and that $A_0(x) \leq A_0(0) \leq \phi(0)$.

Finally we show that $A_1(x) < xA_0(x)$ for all $x > 0$. This follows because

$$\begin{aligned} xA_0(x) - A_1(x) &= x \int_0^\infty (\phi(x+\theta) + \phi(x-\theta))\, \pi(\theta)\, d\theta \\ &\quad - \int_0^\infty ((x+\theta)\phi(x+\theta) + (x-\theta)\phi(x-\theta))\, \pi(\theta)\, d\theta \\ &= -\int_0^\infty (\theta\phi(x+\theta) - \theta\phi(x-\theta))\, \pi(\theta)\, d\theta \\ &= x\, e^{-x^2/2} \int_0^\infty \theta^2 \eta(\theta x)\phi(\theta)\pi(\theta)\, d\theta \geq 0. \end{aligned} \tag{15}$$

The inequality is strict for $x > 0$.

Appendix B: Proofs of the theorems

Proof of Theorem 1: To prove sufficiency, assume that $h(\theta) \to 0$ as $\theta \to \infty$. Then, by l'Hôpital's rule and Assumption (A.4), we have $\omega(\theta) \to 0$ as $\theta \to \infty$. We note, using (10), that for every $M > 0$:

$$\begin{aligned} \int_0^M \theta\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta &\leq -\phi(0) \int_0^M \theta\pi'(\theta)\, d\theta \\ &= \phi(0) \left(-M\pi(M) + \int_0^M \pi(\theta)\, d\theta \right) \leq \phi(0)/2. \end{aligned} \tag{16}$$

Let $\varepsilon > 0$ be arbitrary. Then, using (9) and (13) from Appendix A,

$$\begin{aligned} g(x) = \frac{A_1(x)}{A_0(x)} &= \frac{x \int_0^\infty \theta\eta(\theta x)\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta}{\int_0^\infty (e^{\theta x} + e^{-\theta x})\, \phi(\theta)\pi(\theta)\, d\theta} \\ &\leq \frac{x \int_0^M \theta\eta(\theta x)\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta}{\int_{2M}^\infty e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta} + \frac{\int_M^\infty e^{\theta x}\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta}{\int_M^\infty e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta} \\ &\leq xe^{-2Mx}\eta(Mx) \frac{\int_0^M \theta\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta}{\int_{2M}^\infty \phi(\theta)\pi(\theta)\, d\theta} + \sup_{\theta \geq M} \omega(\theta) \\ &\leq \frac{\phi(0)}{2Me^{Mx} \int_{2M}^\infty \phi(\theta)\pi(\theta)\, d\theta} + \sup_{\theta \geq M} \omega(\theta) < \varepsilon, \end{aligned}$$

using (16), where we have chosen $M = M(\varepsilon) > 0$ such that $\sup_{\theta \geq M} \omega(\theta) < \varepsilon/2$, and then $x > x^* = x^*(\varepsilon, M(\varepsilon))$ with

$$e^{Mx^*} = \max\left(1, \ \frac{\phi(0)}{M\varepsilon \int_{2M}^{\infty} \phi(\theta)\pi(\theta)\, d\theta}\right).$$

This shows that $g(x) \to 0$ as $x \to \infty$.

To prove necessity, assume that it is not true that $h(\theta) \to 0$ as $\theta \to \infty$. Then, $\lim_{\theta\to\infty} \omega(\theta) > 0$. (If $\lim_{\theta\to\infty} \omega(\theta) = 0$, then it follows, again by l'Hôpital's rule, that $\lim_{\theta\to\infty} h(\theta) = 0$, a contradiction.) Hence we have, for some $M > 0$ and $\omega^* > 0$,

$$\inf_{\theta \geq M} \omega(\theta) = \omega^* > 0.$$

This yields

$$\begin{aligned} g(x) &= \frac{\int_0^{\infty} (e^{\theta x} - e^{-\theta x})\omega(\theta)\phi(\theta)\pi(\theta)\, d\theta}{\int_0^{\infty} (e^{\theta x} + e^{-\theta x})\phi(\theta)\pi(\theta)\, d\theta} \\ &\geq \omega^* \frac{\int_M^{\infty} (e^{\theta x} - e^{-\theta x})\phi(\theta)\pi(\theta)\, d\theta}{\int_0^{\infty} (e^{\theta x} + e^{-\theta x})\phi(\theta)\pi(\theta)\, d\theta} \\ &\geq \frac{\omega^*(1 - e^{-2Mx})}{2} \frac{\int_M^{\infty} e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta}{\int_0^{\infty} e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta}. \end{aligned} \tag{17}$$

We now show that the latter expression converges to $\omega^*/2$ as $x \to \infty$. We have

$$\int_0^M e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta \leq e^{Mx} \int_0^M \phi(\theta)\pi(\theta)\, d\theta$$

and

$$\int_M^{\infty} e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta \geq \int_{2M}^{\infty} e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta \geq e^{2Mx} \int_{2M}^{\infty} \phi(\theta)\pi(\theta)\, d\theta.$$

Hence, since M is fixed,

$$\frac{\int_0^M e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta}{\int_M^{\infty} e^{\theta x}\phi(\theta)\pi(\theta)\, d\theta} \to 0 \quad \text{as } x \to \infty. \tag{18}$$

This shows that the ratio of integrals on the right-hand side of (17) converges to one, and hence that the right-hand side of (17) converges to $\omega^*/2$. Hence, $g(x)$ does not converge to zero. This shows that the condition $h(\theta) \to 0$ is necessary and concludes the proof.

Proof of Theorem 2: If $q \geq 1$, then the condition of Theorem 1 is not satisfied, and hence the Subbotin prior is not suitable. Let us assume that $0 < q < 1$. Then, Assumption A holds and, in addition, π is twice differentiable (except at 0) and

$$\frac{\pi'(\theta)}{\pi(\theta)} \to 0, \qquad \frac{\pi''(\theta)}{\pi(\theta)} \to 0,$$

as $\theta \to \infty$. Also, $\pi(0)$ is finite. For such a prior, $x - t(x) \to 0$ as $x \to \infty$ by (the proof of) Theorem 1. Hence, if we can show that $\sigma^2(x) \to 1$ as $x \to \infty$, then the prior is suitable. To show this we consider the function

$$S(\theta, x) := \phi(x + \theta) + \phi(x - \theta),$$

with partial derivatives

$$S'_\theta(\theta, x) := \frac{\partial S(\theta, x)}{\partial \theta} = -(x + \theta)\phi(x + \theta) + (x - \theta)\phi(x - \theta)$$

and

$$S''_{\theta\theta}(\theta, x) := \frac{\partial^2 S(\theta, x)}{\partial \theta^2} = (x + \theta)^2\phi(x + \theta) + (x - \theta)^2\phi(x - \theta) - S(\theta, x).$$

Integrating by parts gives

$$A_2(x) - A_0(x) = \int_0^\infty S''_{\theta\theta}(\theta, x)\pi(\theta)\, d\theta = -\int_0^\infty S'_\theta(\theta, x)\pi'(\theta)\, d\theta.$$

We obtain, for any $M > 0$,

$$\begin{aligned} B(x) := A_2(x) - A_0(x) &= -\int_0^\infty S'_\theta(\theta, x)\pi'(\theta)\, d\theta \\ &= -\int_0^M S'_\theta(\theta, x)\pi'(\theta)\, d\theta + S(M, x)\pi'(M) + \int_M^\infty S(\theta, x)\pi''(\theta)\, d\theta \\ &=: B_1(x) + B_2(x) + B_3(x). \end{aligned}$$

Since

$$\sigma^2(x) = \frac{A_2(x)}{A_0(x)} - \left(\frac{A_1(x)}{A_0(x)}\right)^2 = 1 + \frac{B(x)}{A_0(x)} - \left(\frac{A_1(x)}{A_0(x)}\right)^2$$

and $A_1(x)/A_0(x) \to 0$ by Theorem 1, it suffices to show that $B(x)/A_0(x) \to 0$.

Let $\varepsilon > 0$. Then, using (11), there exists an $M = M(\varepsilon) > 0$ such that

$$\frac{B_3(x)}{A_0(x)} \leq \frac{\int_M^\infty S(\theta, x)\, \pi''(\theta)\, d\theta}{\int_M^\infty S(\theta, x)\, \pi(\theta)\, d\theta} \leq \sup_{\theta \geq M} \frac{\pi''(\theta)}{\pi(\theta)} < \varepsilon/2.$$

Also, for $x > M + 1$,

$$|B_1(x)| \leq \int_0^M \phi'(\theta - x)|\pi'(\theta)|\, d\theta \leq \phi'(M - x) \int_0^M |\pi'(\theta)|\, d\theta$$
$$\leq \phi'(M - x)\pi(0) = (x - M)e^{-x^2/2}e^{Mx}\pi(0)\phi(M),$$

$$|B_2(x)| \leq 2\phi(x - M)\,|\pi'(M)| = 2e^{-x^2/2}e^{Mx}|\pi'(M)|\phi(M),$$

and

$$A_0(x) \geq \int_{2M}^{\infty} S(\theta, x)\,\pi(\theta)\, d\theta \geq e^{-x^2/2}e^{2Mx} \int_{2M}^{\infty} \phi(\theta)\pi(\theta)\, d\theta.$$

Hence, for large enough x,

$$\frac{|B_1(x) + B_2(x)|}{A_0(x)} \leq \frac{e^{-Mx}\phi(M)\,((x - M)\pi(0) + 2|\pi'(M)|)}{\int_{2M}^{\infty} \phi(\theta)\pi(\theta)\, d\theta} < \varepsilon/2.$$

Thus we have proved that $|B(x)|/A_0(x) < \varepsilon$. Hence, $B(x)/A_0(x) \to 0$, which shows that $\sigma^2(x) \to 1$ as $x \to \infty$, and concludes the proof.

References

Bauwens, L., Lubrano, M., Richard, J.-F., 1999. Bayesian Inference in Dynamic Econometric Models. Oxford University Press, Oxford.

Bayes T., 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53, 370-418.

Bernardo, J.M., 1979. Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society B, 41, 113–147 (with discussion).

Bickel, P.J., 1983. Minimax estimation of the mean of a normal distribution subject to doing well at a point. In Rizvi, M.H., Rustagi, J.S., Siegmund, D., editors, Recent Advances in Statistics: Papers in honor of Herman Chernoff on his sixtieth birthday. Academic Press, New York, pp. 511–528.

Bickel, P.J., 1984. Parametric robustness: Small biases can be worthwhile. The Annals of Statistics, 12, 864–879.

Danilov, D., Magnus, J.R., 2004. On the harm that ignoring pretesting can cause. Journal of Econometrics, 122, 27–46.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348–1360.

Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics, 35, 109–135.

Hans, C., 2009. Bayesian lasso regression. Biometrika, 96, 835–845.

Jeffreys, H., 1961. The Theory of Probability (3rd edition). Clarendon Press, Oxford.

Laplace, P.-S., 1814. Essai Philosophique sur les Probabilités. Cambridge Library Collection—Mathematics, Fifth edition (2009), Cambridge University Press, Cambridge.

Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York.

Magnus, J.R., 2002. Estimation of the mean of a univariate normal distribution with known variance. Econometrics Journal, 5, 225–236.

Magnus, J.R., Durbin, J., 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. Econometrica, 67, 639–643.

Magnus, J.R., Powell, O., Prüfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. Journal of Econometrics, 154, 139–153.

Park, T., Casella, G., 2008. The Bayesian lasso. Journal of the American Statistical Association, 103, 681–686.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociological Methodology, 25, 111–163.

Schuster, E.F., 1984. Classification of probability laws by tail behavior. Journal of the American Statistical Association, 79, 936–939.

Subbotin, M.Th., 1923. On the law of frequency of error. Mathematicheskii Sbornik, 31, 296–301.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (Methodological), 58, 267–288.